27 September 1974, Volume 185, Number 4157 SCIENCE

Classification: Purposes, Principles, Progress, Prospects

Clustering and other new techniques have changed classificatory principles and practice in many sciences.

Robert R. Sokal

The origin of the science of classification goes back to the writings of the ancient Greeks (*1*). However, the process of classification, the recognition of similarities, and the grouping of organisms and objects based thereon dates back to primitive man (*2, 3*). Even before the advent of man, classificatory ability must have been a component of fitness in biological evolution. Regardless of whether behavior is learned or instinctive, organisms must be able to perceive similarities in stimuli for survival (*4*). Thus the recognition of similarities in patterns of sensory input is probably as old as the earliest forms of sense perception in living organisms.

The study of classification has always had two major interrelated components: How do we classify? and How should we classify? If we restrict the discussion to classification by man of the world around him, the first component falls into the domain of the psychology and philosophy of sense perception: What is similarity? How do human beings recognize similarity? What are the criteria, conscious and unconscious, by which man groups objects and events into some system? What are the relationships between the names man gives to classes of entities and their objective definition? Are there individual differences in the perception of similarity and in the ability to group objects, and how do these affect everyday and scientific communication? The second component is the subject matter of taxonomy, the science of classification. It arises from the conceptualization of the questions raised earlier. Knowing how man classifies, can one derive general principles of classification from this knowledge and, if so, are these principles the best conceivable ones, given that choices between different classificatory principles and procedures are permitted? This article will mainly concern itself with the second aspect of classification, the principles and procedures.

Classification is an important aspect of most sciences, and similar principles and procedures have been developed independently in various fields. Fortunately, substantial interdisciplinary communication has accompanied recent progress (*5–7*), and a body of general classificatory theory and methodology is rapidly being developed, a task that is attracting the interest of statisticians and mathematicians (*8*). In classification, theory has frequently followed methodology and has been an attempt to formalize and justify the classificatory activity of workers in various sciences. In other instances, classificatory systems have been set up on a priori logical or philosophical grounds and the methodology tailored subsequently to fit the principles. Both approaches have their advantages and drawbacks; modern work tends to reflect an interactive phase in which first one and then the other approach is used, but in which neither principles nor methodology necessarily dominate.

Computers and Classification

Even though the principles and many of the mathematical ideas underlying modern classificatory methods antedate the appearance of the electronic computer, the recent great increase in classificatory work is intimately related to the development of this new tool. It is difficult to assess the degree to which the acceleration of work in classification is due to the almost simultaneous rapid increase in the availability and capability of computers. It might be argued that other developments in modern science could have renewed the interest in classification. An unprecedented number of scientists is at work and new, automated methodologies are yielding information on many properties of numerous objects. This process alone could easily have forced the development of new classificatory techniques to cope with the flood of information. Yet without computers it is inconceivable how such developments could have been implemented.

Computers play a central role in modern classification for four reasons. First, they have helped to find solutions to problems that were analytically intractable (or at least seemed so to the scientists involved). The numerous algorithms for minimum length problems or iterative reallocation procedures (*9*) are cases in point. Second, computers are able to carry out computations whose numerical solutions were known but exceedingly tedious, for example, the computation of eigenvalues of large matrices. Also, com-

The author is professor of biological sciences at the State University of New York at Stony Brook, Stony Brook 11790. This article is adapted from the Presidential Address delivered to the Classification Society at the University of Chicago, 24 April 1972.

puters have been able to classify simultaneously far larger numbers of objects, using many more features of these objects, than any human taxonomist. A third and most important by-product of the application of computers has been the necessary development of algorithms for classification, which in turn has led to attempts to objectivize and optimize the classificatory process. This is a clear example of the influence of methodological development on theory. Finally, because of the general development of pattern recognition and perceptron technology, the availability of computers has given rise to fundamental investigations into how human beings and other organisms perceive the world around them. One hopes, in these studies, to have the computer imitate man or other organisms in their perceptive and classificatory abilities. Such work relates largely to the first of the questions raised earlier (How do we classify?) but it has also brought in its wake debate on whether the classificatory systems to be established in various sciences should conform to man's intuitive "natural" classificatory ability or whether other criteria of goodness of a classificatory system are preferable. Also, computer taxonomy permits objective checks and experiments of various hypotheses concerning the science of sense perception—a relatively undeveloped field so far.

Definition of Terms

Before we proceed we must guard against possible confusion: several important terms are employed with varying meanings in different sciences (*10*).

Classification will here be defined as the ordering or arrangement of objects into groups or sets on the basis of their relationships. These relationships can be based on observable or inferred properties. Some philosophers, mathematicians, and statisticians also employ the term "classification" for what is here called "identification," the term being defined as the allocation or assignment of additional unidentified objects to the correct class, once such classes have been established by prior classification. Thus we "identify" an object as being a chair, or a plant as being a buttercup.

In addition to indicating a process, the term classification is frequently employed to denote the end product of this process. Thus the result of classification is a classification. It seems better to term such an end result a "classificatory system" (*7*).

The term "taxonomy" is used here to mean the theoretical study of classification including its bases, principles, procedures, and rules (*11*). This would include classification as well as identification. It is the science of how to classify and identify.

The term "taxon" (plural: taxa) is useful to designate a set of objects of any rank recognized as a group in a classificatory system. A taxon has therefore been arrived at by some classificatory procedure, but not necessarily as the result of an explicit methodology; for example, one can have taxa in a folk taxonomy.

Purposes of Classification

Much classificatory work in various sciences aims to describe what is known as the "natural system." This is a difficult concept and differs in meaning across disciplines and even among workers in any one field. Natural systems are believed to be in accord with nature. If it is the purpose of science to discover the true nature of things then it is the purpose of a correct classification to describe objects in such a way that their "true" relationships are displayed. In many sciences this has led to essentialist systems whose philosophical origins go back to Aristotle. The difficulty with essentialism is that it is based on Aristotelian logic expressed in axioms that give rise to properties that are inevitable consequences of these axioms. Such conditions apply to classifications of some entities, such as colors or geometrical figures, but not to others.

One view of natural classifications is that they reflect the natural processes that have led to the observed arrangement of the objects. One hopes from such an ordering to learn about the laws governing the behavior of these objects. In biology there supposedly is such a natural system, reflecting the end products of the evolutionary process. Yet natural systems are not necessarily isomorphic with the common, mutually exclusive classificatory arrangements employed by taxonomists in various fields. The process giving rise to the differentiated objects may be such as to create overlapping classes or fuzzy boundaries because of anastomosing processes. The borrowing of stylistic features in human artifacts is an example in point. Others would be textual materials, hybrid languages, social systems, or organisms.

All classifications aim to achieve economy of memory. The world is full of single cases: single individuals of animal or plant species, single case histories of disease, single books, rocks, or industrial concerns. By grouping numerous individual objects into a taxon the description of the taxon subsumes the individual descriptions of the objects contained within it. By saying that Jean Duval speaks French, we imply that his linguistic inventory resembles that of millions of other persons in the taxon "French-speaking persons," and we save ourselves a whole catalog of statements about the particular word lists and sentence structures familiar to Duval. Unless we qualify our statement further we are lumping together varieties of thought, speech, and writing patterns collectively known as the French language, and without a clearer definition of boundaries we cannot be certain whether local dialects such as Parisian argot or Provençal are included, or which variety it is that Duval speaks. Yet without the ability to summarize information and attach a convenient label to it we would be unable to communicate.

Yet another purpose of classification is ease of manipulation. The objects are arranged in systems (that may or may not be hierarchic) in which the several taxa can be easily named and related to each other. If the relationships are very complex, as are functional roles of individuals in certain societies, for example, no easy labeling or handling of the taxa will be possible. Ease of retrieval of information from a classificatory system is also a criterion frequently considered desirable.

The paramount purpose of a classification is to describe the structure and relationship of the constituent objects to each other and to similar objects, and to simplify these relationships in such a way that general statements can be made about classes of objects. The definition, description, and simplification of taxonomic structure is a challenging task. It is easy to perceive structure when it is obvious and discontinuous. Disjoint clusters separated by large empty regions are unambiguous. Thus horseshoe crabs or ginkgo trees are unique species quite different from their nearest relatives. A language such as Basque is in a similar position. But this situation is not typical. Much

of what we observe in nature changes continuously in one or another characteristic, but not necessarily with equally steep gradients for each characteristic. Where should boundaries be drawn in such cases? Must classification be a drawing of boundaries? Would an adequate description and summarization of the continuity of the objects be preferable to artificially erected boundaries? Uniform continuous change is, of course, not very frequent in nature. Centripetal forces frequently hold together a certain structure over a given domain and loosen their control only at zones of rapid intergradation. In biology, stabilizing selection within a gene pool would be a case in point. Another example of such a normative force is the effect on regional languages or dialects of the publication of newspapers and the broadcasting of radio and television programs. Thus the boundary between Catalan and Castilian in Spain is undoubtedly reinforced by the existence of media and centers for diffusion of the respective languages and cultures. Often the clusters will obey gravitation-like laws with boundaries definable as equilibrium points, while in other cases diffusion or stepping-stone models best describe the transition between clusters. Taxonomists must decide the relative importance of diameters and densities of clusters, the number of objects, and the gaps between the clusters.

Classifications that describe relationships among objects in nature should generate hypotheses. In fact the principal scientific justification for establishing classifications is that they are heuristic (in the traditional meaning of this term as "stimulating interest as a means of furthering investigation") and that they lead to the stating of a hypothesis which can then be tested. A classification raises the question of how the perceived order has arisen, and in a system in which forces and relationships are transitory one may conjecture about the maintenance of the structure. Examples are inferences about evolutionary lineages obtained from biological classifications based on morphological or biochemical characters, inferences about population structure in biology and anthropology resulting from patterns of geographic variation, and inferences about acculturation which certain models of linguistic and artifactual evolution engender in anthropology.

The search for immanent structure in nature is far from the only purpose of classification. Especially in applied, practical fields the question is often asked: What is the best classification of the objects at hand into two or three or k classes? In regionalization studies a given political area is to be divided into a fixed number of districts given some criterion of optimality. What is the best way to subdivide a county into five voting districts to achieve maximal—or in some recent redistricting problems minimal—intraclass homogeneity? Many routing problems can be considered classification problems in this sense. If a bakery possesses four trucks how can it best route these through the city to cover the set of n grocery stores in the city at minimal cost or in minimal time?

The two kinds of approaches, the search for natural structure and the imposition of an external constraint by fitting the data to a fixed number of classes, are not necessarily categorically distinct. I suspect that many biological taxonomists, without explicitly saying so, assume that they already know the major subdivisions of the organisms they study and only need to allocate properly the finer taxonomic units to these major subdivisions. In biological and some other classifications it is sometimes stated that the number of major subdivisions should be partly a function of the number of included taxonomic units. Such a scheme clearly is not based on fundamental scientific principles but largely on considerations of practicality. They may also be related to the number of names human beings are able to recall from a data base (*3*).

Principles of Classification

Of the various principles applied in recent classificatory theory, the distinction between monothetic and polythetic classification, first clearly enunciated by Beckner (*12*), is probably of greatest importance. Monothetic classifications are those in which the classes established differ by at least one property which is uniform among the members of each class (*13*). Such classifications are especially useful in setting up taxonomic keys and certain types of reference and filing systems. From the practical point of view of information retrieval it is obviously desirable that certain properties of taxa be invariant (*14*).

In polythetic classifications, taxa are groups of individuals or objects that share a large proportion of their properties but do not necessarily agree in any one property. Adoption of polythetic principles of classification negates the concept of an essence or type of any taxon. No single uniform property is required for the definition of a given group nor will any combination of characteristics necessarily define it. This somewhat disturbing concept is readily apparent when almost any class of objects is examined. Thus it is extremely difficult to define class attributes for such taxa as cows or chairs. Although cows can be described as animals with four legs that give milk, a cow that only has three legs and does not give milk will still be recognized as a cow. Conversely there are other animals with four legs that give milk that are not cows (*15*). It is similarly difficult to find necessary properties of the class "chairs." Properties that might commonly be found in any chair may be missing in any given piece of furniture that would clearly be recognized as a chair. These somewhat contrived examples can be bolstered by numerous instances of classification ranging from archeology to zoology (*16*). When viewed from a historical perspective we find remarkable parallels in the gradual rejection of the type concept and the adoption of polythetic criteria in these various disciplines.

A corollary of polythetic classification is the requirement that many properties (characters) be used to classify objects. This is true of almost any type of object being classified. Biological organisms, with their complex sequences of nucleotides and great diversity of structure and function, are rich in variability and yield numerous characters; but artifacts or art objects, languages, industries, or case histories of physical or mental disease yield many characters as well. Some may argue that a few attributes are sufficient to characterize taxa in these fields. Most instances quoted to support this point of view are cases of identification. Once a classification has been established, few characters are generally necessary to allocate objects to the proper taxa. But it is unlikely that few characters will suffice to establish the taxa in the first place. Initial classifications based on few characters usually have had to be modified once information on additional characteristics was acquired. Diseases not differentiated in earlier times now represent separate clinical entities with the ac-

cumulation of new knowledge; Linnaeus's *Vermes* has become numerous animal phyla.

Classifications based on many properties will be general; they are unlikely to be optimal for any single purpose, but might be useful for a great variety of purposes. By contrast, a classification based on few properties might be optimal with respect to these characters, but would be unlikely to be of general use (*17*). Thus an alphabetical ordering of books by author in a library will be the ideal classification for an alphabetical author catalog but will not contribute to a meaningful classification by subject matter. A classification of plants by growth form will not reflect the natural taxa, although it might be useful from an ecological or landscaping point of view. For many practical purposes special classifications based on few characters are desirable.

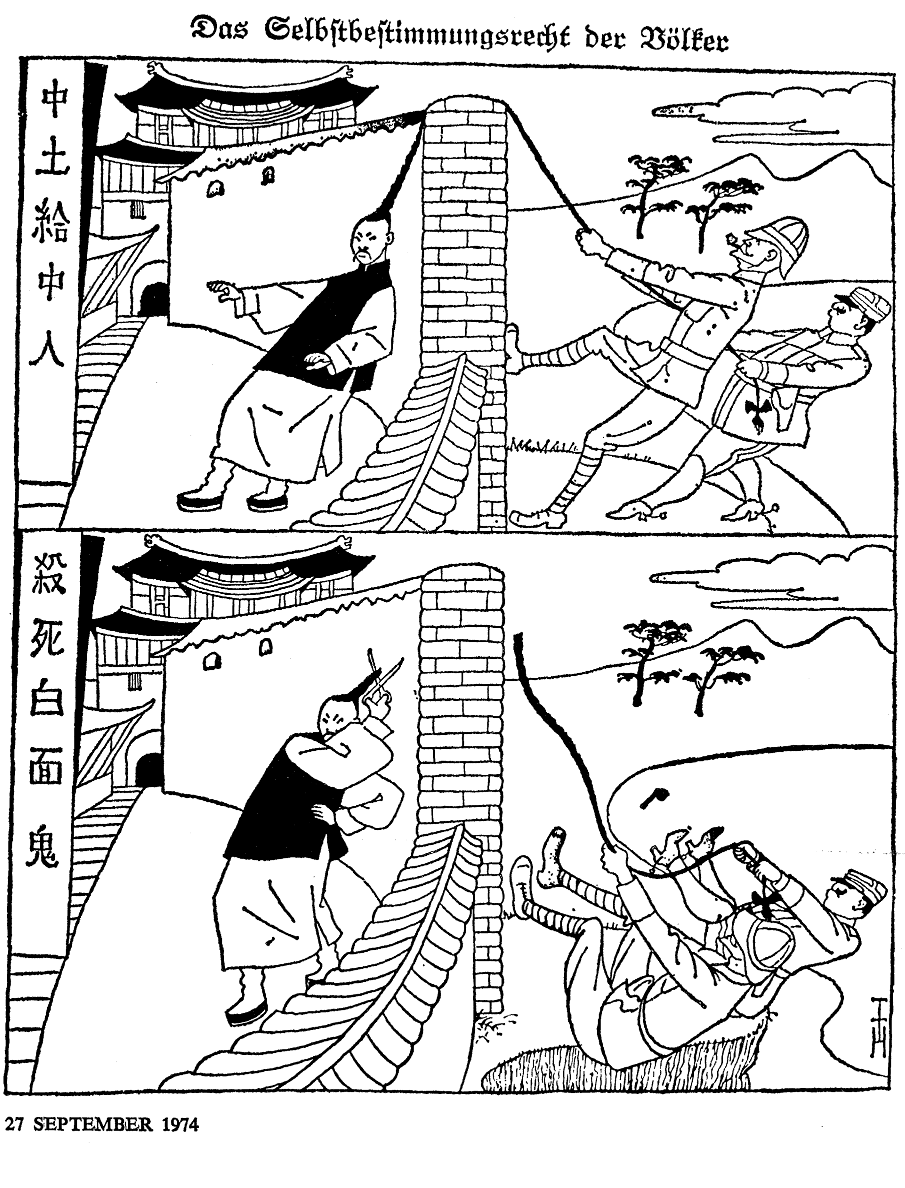
Weighting. The problem of weighting characters has troubled taxonomists in all disciplines. Should certain characters be weighted more heavily than others? Many biologists maintain that traits indicating common evolutionary descent be weighted more heavily than others, and they weight the discordant characters less than others when constructing a classification. Established differentiae between diseases, languages, or cultures might be similarly emphasized. The difficulty with such weighting is that one needs initial classifications to provide weights for the characters. But once classifications are correct there is little value in computing weights for the characters that established them, except for future identification of unknown objects. Many modern taxonomists have therefore adopted the doctrine of equal weighting of characters for classification. Those that do not advocate equal weighting have been forced to state a basis upon which they propose to weight characters—and many such proposals are found wanting (*18*).

Once a taxonomist becomes convinced that a particular trait is of great importance in dividing up his material he subsequently, almost inevitably, becomes quite selective about the other evidence he collects. He will more readily use characters that support his earlier views and weight fewer characters that are discordant. Such tendencies are found in every field of classification. For example, having decided that locusts are divisible into migratory and solitary phases on the basis of body proportion, entomologists attempted to fit discoveries of other differences in

Figs. 1 and 2. Cultural biases in depicting racial differences. Fig. 1 (above). The drawings show English and Chinese representatives during the Opium War signing a treaty. The contemporary Chinese artist (*46*) emphasized the prominent noses of the Westerners by receding the upper lip. The curly hair of the "Western barbarians" was equally stressed. These are not wigs, since other illustrations in the same source show every Englishman, including fighting and looting British Army privates, with similar exaggerated curly hair (see cover). Curiously, the hairiness of the Westerners or their tallness are not stressed in this or other illustrations of the Opium War, although most present-day Chinese when queried will stress these features of Europeans as important differentiae. Note also the black figures on the cover illustration. These are not Africans, but quite likely were Indians, possibly Sikhs. They were depicted as far darker than Sikhs are actually. It is interesting that the acculturation is evident in modern Chinese cartoons where characters intended to be Chinese frequently appear undistinguishable from Europeans, at least to the Western observer. Fig. 2 (facing page). This illustration, drawn by a well-known German cartoonist in 1925 (*47*), stresses prominent check bones and slanted eyes. [Reproduced by courtesy of Scherz, Bern (*47*)]

pigmentation and behavior into the framework set up by the original classification into phases. It took considerable conceptual liberation from the earlier system to arrive at the more complicated, essentially polythetic view of phase formation in locusts (*19*). Likewise, after a dimorphism was observed in the shapes of aphid galls, a detailed analysis (*20*) revealed that there was also an important dimorphism in the life histories of the insects. Continued attention to the superficially more impressive gall shape characters led to misclassification of some of the galls.

The selection and recognition of characters is but an extreme example of weighting. Traits that are not employed are given a weight of zero. Cultural and personal biases affect character selection in virtually every field of classification. To take an example from physical anthropology, let us compare two human populations belonging to different races. Samples of natives from, say, England and China, would differ in numerous traits. It is possibly a truism that each and every member from a sample of one group could, with statistical assurance, be distinguished from every member in a sample of the other group, if characters were selected in an unbiased fashion. These characters would undoubtedly cover aspects of external appearance, musculature, skin color, pubescence, and bone measurements, for example. Yet if members of these two groups were asked to describe each other, the differential characters noted by them would be quite different. Thus, when a typical European or American is asked to list the salient distinguishing traits of the Chinese he will most frequently mention the "slanted eyes" and secondly the yellow skin color. Prominent cheekbones and straight black hair would be mentioned frequently as well. Possibly because eye shape and the presence of the epicanthic fold are quite variable among the Chinese, they do not describe Europeans by such characters. The descriptors that readily come to the minds of the Chinese are the tallness of the Westerners, their blond or brown curly hair that is quite absent in China, their hairiness, as well as their prominent noses. These culturally conditioned differentiae not only lead to epithets for the other race in both cultures but some are also seen in drawings by Chinese and European artists (see Figs. 1 and 2). The two populations can be compared on either the European or the Chinese set of characters but they clearly differ on both of these and in other properties as well. Although popular descriptive terms may not affect the more objective judgment of physical anthropologists, differences of this sort may still guide classificatory behavior in various covert ways. A well-known analogous case is the great diversity of the negroid populations of Africa which, to the casual observer, is hidden by the blackness of the skin.



Not only the culturally conditioned biases of entire populations need to be corrected by objective classificatory systems, but individual variability in the perception of similarity and shape must be allowed for. It is now well established that individual differences in recognition of form and shape are in part due to individual differences in eye scan patterns (*21*). Attempts are being made to show individual components of taxonomic judgment (*22*). While there is considerable commonality in judgment of similarities, individual observers do differ in the importance which they intuitively assign to different aspects of shape, form, or color, for example (see Fig. 3).

Progress in Classification

A convenient way of developing classifications is to compute functions that yield similarities or dissimilarities (distances) between all objects taken a pair at a time. A symmetric matrix of such similarity or dissimilarity coefficients is then analyzed to represent their relationships as clusters or in various other ways. The type of pair function will depend on the data to be analyzed. Binary data usually lead to association coefficients [section 4.4 in (*6*)], continuous variables to some type of distance or correlation coefficient.

Much recent progress in classification has consisted of devising methods of clustering. This would suggest that

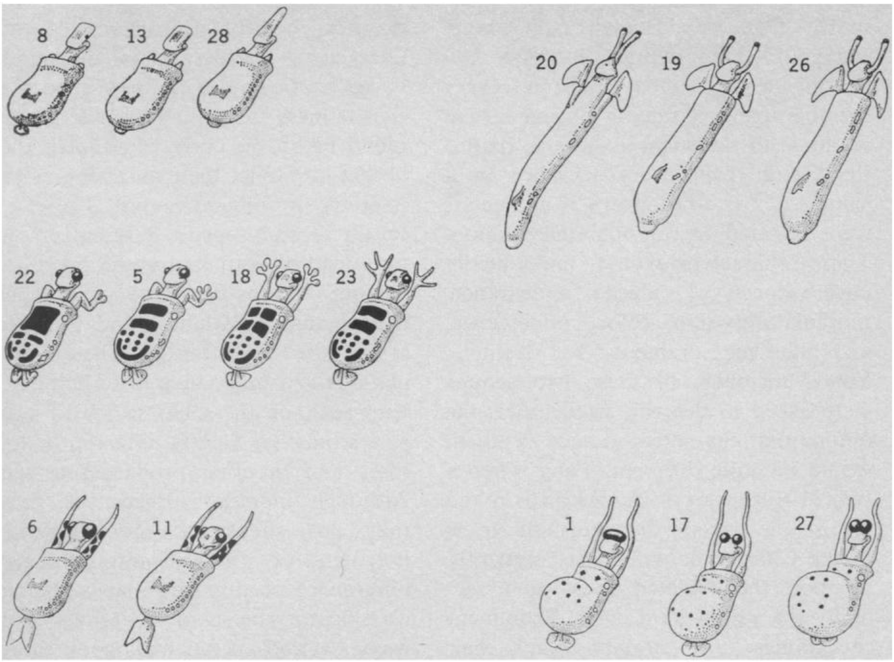


Fig. 3. Caminalcules, imaginary animals created by J. H. Camin, serve to illustrate individual differences in taxonomic judgment. Twenty-nine different organisms were presented to a large number of persons ranging from professional taxonomists to children. The data presented here are a small sampling from the study in progress (*48*). Three taxonomists, A (a distinguished systematic entomologist), B (an invertebrate paleontologist), and C (a graduate student in paleontology), were asked to group the organisms by their similarities. From the classifications established by the three persons, the following relationships illustrated by groups of Caminalcules in the figure can be extracted. Taxonomists A and C thought 13 was more similar to 8, but B placed it closer to 28. All three taxonomists thought 6 was most similar to 11. While taxonomist C placed 5 and 18 together, taxonomist A grouped 5 with 22, and 18 with 23, and B did not form a close group with any of these Caminalcules. Taxonomist A thought 17 was most similar to 1, C held it most similar to 27, and B described the three organisms as equally similar. Taxonomists A and C recorded 19 most similar to 26, but B considered it closer to 20. By multiple regression of the similarities implied by the taxonomists on 112 objectively defined criteria differentiating the 29 animals (these criteria were not furnished to the experimental subjects), the relative importance of various criteria in judging taxonomic similarity can be inferred. The judgments by persons A and C were more similar to each other than either was to B; most dissimilar were B and C. Table 1 shows which features of the organisms appeared important to each of the three taxonomists. A plus sign indicates a feature important to the stated taxonomist. No one feature was important to all three persons, and quite different aspects of the creatures were stressed by the subjects.

the concept of a cluster is clearly understood by those who do the clustering. Regrettably this is not always so. The various algorithms developed for clustering impose a structure on the objects to be clustered (generally known as OTU's, operational taxonomic units) which are represented as vectors of descriptors (character states). An attempt at clustering the OTU's implies a belief that they exhibit some structure and are not randomly or uniformly distributed through the hyperspace defined by the descriptors. Definitions of clusters are hard to come by. One book on cluster analysis does not define clusters at all (*23*), another (*6*) deliberately defines it loosely as "sets of OTU's in hyperspace that exhibit neither random nor regular distribution patterns and that meet one or more of various criteria imposed by a particular cluster definition." A more intuitively appealing but at the same time more restrictive definition is "a set of objects characterized by the properties of isolation and coherence" (*7*). Clusters can be described by the different densities encountered on sweeping out the hyperspace. Properties of clusters include their location in space (some measure of central tendency), their dispersion, their shapes (for example, hyperspheres or hyperellipsoids), their connectivity (a measure of how many of the pairs of OTU's within a cluster are more similar to each other than a certain arbitrary criterion), and the magnitude of gaps between clusters.

Clustering algorithms can be agglomerative or divisive. In agglomeration the OTU's can be considered as the disjoint partition of the whole set and can be aggregated to form ever larger

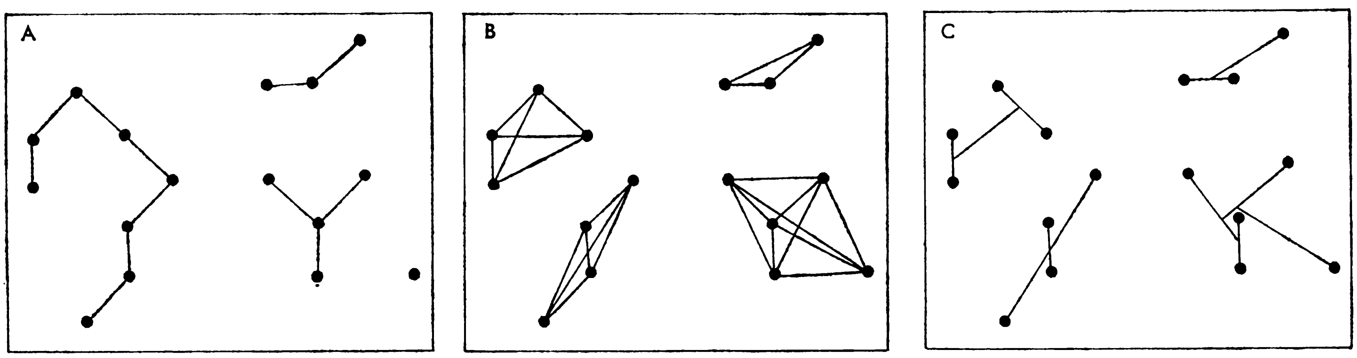


Fig. 4. The effect of different clustering methods. Sixteen arbitrary points plotted in a two-dimensional space are clustered by (A) single linkage clustering, (B) complete linkage clustering, and (C) an average linkage clustering method (unweighted centroid clustering). Approximately corresponding stages in the clustering process are shown for the three methods yielding four open, loose clusters in single linkage (one cluster has a single member), four tight and discrete clusters in complete linkage and an intermediate solution by an average linkage method. Continuation of the clustering process in all three cases would yield the conjoint partition in which all 16 points would form a single cluster. [Illustrations adapted from (*6*)]

clusters until the conjoint partition, an entire set consisting of all the OTU's, is reached. The converse, divisive approach is to break down the conjoint partition into subsets until the disjoint partition is reached where each subset is a single OTU. In most clustering techniques, especially in polythetic methods, the agglomerative approach is preferred for practical reasons in devising a workable computer algorithm.

The following three agglomerative clustering techniques are employed frequently. (i) In the single linkage method candidates for membership will join a cluster if they are connected to any member of the cluster by a single pairwise relation at or above the accepted criterion of similarity. This algorithm tends to define long strung-out sets of points as clusters and will group even fairly evenly distributed OTU's into straggly appearing clusters. Single linkage clustering has two desirable mathematical properties. It is invariant under monotone transformations of the similarity coefficients and is isomorphic with the "shortest spanning tree" (as discussed below) covering the same set of OTU's. Yet the imposition of straggly relationships on OTU's does not generally yield acceptable classifications. (ii) The converse approach is complete linkage clustering in which a candidate for membership in a cluster must connect to all present members by the accepted criterion before being permitted to join. Complete linkage clustering, also monotone-invariant, results in tight ball-shaped clusters. (iii) To overcome the extremes of these two approaches, the commonly employed average linkage clustering techniques compute the similarity of a candidate OTU with an established cluster as some average of the similarity of that OTU with all the members of the cluster (see Fig. 4). The ideal clustering method would be adaptive, in that the clustering algorithm would modify itself based on the most likely hypothesis of constellation of points in the hyperspace. It would first "feel out" this constellation and, on finding it to be a certain type of structure, would then change the algorithm to reinforce the hypothesized structure.

The results of cluster analysis are often represented by dendrograms which are hierarchic representations of the similarity relations among the OTU's. An example is shown in Fig. 5. One axis is graduated in the similarity or dissimilarity scale and the branching points along the scale indicate the resemblance between the stems being joined.

Table 1. Features of Caminalcules that appeared important to three taxonomists.

Features of Caminalcules	Taxonomists A	Taxonomists B	Taxonomists C
Horns on head		+	
Stalked eyes	+		+
Groove in neck		+	
Anterior appendage			
Length	+		+
Flexion		+	
Subdivision	+		
Bulb		+	
Posterior appendage			
Disklike	+		+
Platelike	+		
Anterior abdomen spots			+
Posterior abdomen bars	+	+	
Abdomen			
Width	+		
Large pores	+		+
Small pores		+	

Classifications need not be hierarchic and the clusters may overlap (intersect). The whole idea of hierarchic, nonoverlapping (mutually exclusive) classifications which is so attractive to the human mind is currently undergoing reexamination. From studies in a variety of fields the representation of taxonomic structure as overlapping clusters or as ordinations appears far preferable. By ordination we mean projection of the OTU's in a space of fewer dimensions than the original number of descriptors. When tested by any of several measures of distortion, ordinations in as few as two or three dimensions will frequently represent the original similarity matrices considerably more faithfully than dendrograms. A common means of ordination is by principal components analysis (*6, 24*) or any of its variants, which define a space of $k<n$ dimensions which are linear combinations of the n original characters describing the OTU's (see Fig. 6). An elegant method popular in recent years is nonmetric multidimensional scaling (*6, 25*) which ordinates the OTU's into a space of predetermined dimensionality on a criterion of best fit to a monotonic function of the original similarity matrix. Since only rank order and monotonicity of the similarity coefficients is assumed, this method is particularly robust and gives unusually good results in a great variety of fields.

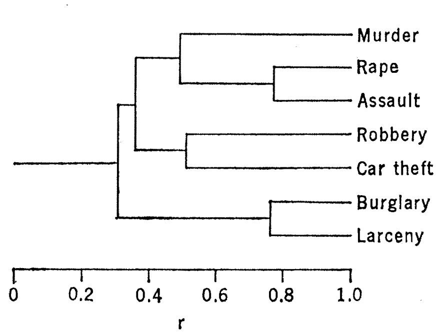


Fig. 5. Results of a cluster analysis of seven crimes based on their incidence in 16 U.S. cities. Relationships are represented by means of a dendrogram (phenogram) in which the implied similarity between any two crimes can be read off the horizontal axis graduated in correlation coefficient scale. Thus burglary and larceny resemble each other at a level of 0.76 while the implied correlation between murder and larceny is only 0.31. The phenogram shows a clear separation between crimes of property and the more serious crimes, and distinguishes a cluster of crimes of violence (*49*).

Ordination obviously will give especially good results where clear clusters are lacking and the OTU's are distributed in continua such as streaks through hyperdimensional space. In fact the development of ordination in the classification of plant communities has led to a fundamentally different concept of plant sociology, contrasting the older notion of discrete communities classifiable in a hierarchic nonoverlapping manner with the continuum concept in which members of associations (species) are replaced on a continual, partially overlapping basis yielding polythetic taxa (*26*).

When connectivity of the OTU's is considered we turn naturally to graph theory and the notions of graphs and trees (*27*). Clusters are partially or fully connected graphs that depend on the degree of similarity required for admission of candidates to the cluster. Connectivity provides us with measures of the tightness of clusters. An important family of graphs are "trees," which are minimally connected graphs—that is, those that contain only one direct or indirect path between every pair of OTU's in a cluster. When length is given to these paths, it becomes possible to find the shortest tree, known as the "shortest spanning tree." This tree is equivalent to finding those linkages among OTU's that would be found by single linkage clustering. When shortest spanning trees in the full (n-) dimensional space are superimposed on the OTU's in a two- or three-dimensional ordination they help to indicate major distortions since linkages between nonadjacent OTU's suggest similarities in other regions of the

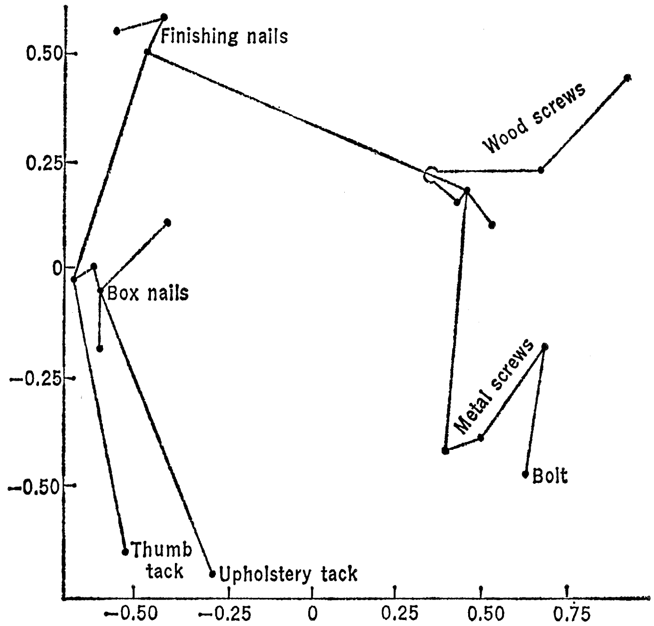


Fig. 6. An ordination of 19 metal fasteners based on 23 characteristics. The graph shows the first two principal axes, which are linear combinations of the 23 descriptive characters. The first axis (abscissa) separates objects that are turned (screws and bolts) from objects that pierce the substrate (nails and tacks), and reflects also differences in the surfaces of the heads and in the points of the shafts. The second axis (ordinate) is affected by the relative magnitude of head versus stem and by the angle of head to the shaft. The lines connecting the points give the minimum spanning tree for the 19 objects in a full-dimensional space and serve as an indicator of possible distortion. Objects close to each other in the two-dimensional space but not directly connected are likely to be differentiated in a dimension not shown here. An example is the apparent difference between the thumbtack and the upholstery tack, which in the full-dimensional space resemble separate box nails more than they do each other. The overall classification achieved by this method is quite satisfactory, although the correspondence between the similarities ordinated in the three-dimensional space and the original similarity matrix is not as great as in other studies.

n-dimensional space not represented in the ordination (see Fig. 6).

A major application of minimal length trees in recent years has been in the reconstruction of evolutionary sequences in which some criterion of parsimony of evolution is used (see Fig. 7). Such work has been done with the amino acid sequences in proteins of different species (*28*) and with OTU's based on a variety of morphological, gene frequency, or biochemical characters (*29*). Some of the methods involve so-called Wagner trees which are directed Steiner trees of minimal length in which hypothetical points are permitted to be constructed in order to shorten the length of the overall tree. The development of operational methods of evolutionary analysis have contributed greatly to the development of a new synthesis in taxonomic theory between the pheneticists who wish to erect natural systems not on evolutionary descent but on resemblances based on a multiplicity of characters (*30*) and the cladists who wish to consider recency of branching sequences—that is, recency of common ancestry exclusively (*31*). The application of methods for reconstructing branching sequences to other fields such as languages, artifacts, and migration patterns appears to hold considerable promise.

Prospects for Classification

The generality of the applications discussed above rests on the near universal desire in many fields of science to classify OTU's into taxa and on the wide availability of multiple descriptors for these OTU's. This almost necessitates computer handling in a multivariate manner as I have described. The problem of correct classification of phenomena is one that constantly crops up in scientific work, and many controversies could be avoided, or at least the area of disagreement narrowed and the point under contention refined, if proper principles of classification were adhered to. An example in point is the recent controversy regarding the classification of psychotic symptoms (*32*). The areas of disagreement could have been narrowed if the existence and desirability of polythetic taxa had been made explicit by the writers.

The extent of the application of these methods has been far-ranging. Computer classification has been applied to animals (*33*), plants (*34*), microorganisms (*35*), vegetation (*36*), soils (*37*), physical and psychiatric diseases (*38*), physical anthropology (*39*), languages and cultures (*40*), and stream sediments (*41*). Other work includes economics (*42*), market research (*43*), naval studies (*44*), and classifications of manuscripts in the humanities (*45*). Among the major unsolved problems in computer classification are (i) the appropriate coding of characters so as to give unbiased measures of similarity among objects to be clustered; (ii) the best criteria of optimality for clustering or ordination to represent the similarity among objects with a minimum of distortion; and (iii) tests of significance of taxonomic structure found in nature.

Various instruments or data-gathering devices that will automatically yield large data sets are becoming frequent in a variety of fields. Autoanalyzers, amino acid sequences, radar scanners, and scanning electron microscopes are examples of such automated sensing devices whose output will not be digestible unless we treat them by various methods of computer classification. The use of electronic data processing for the description of the OTU's, their arrangement into taxa, the preparation of geographic distribution maps, and for keys and other identifying devices

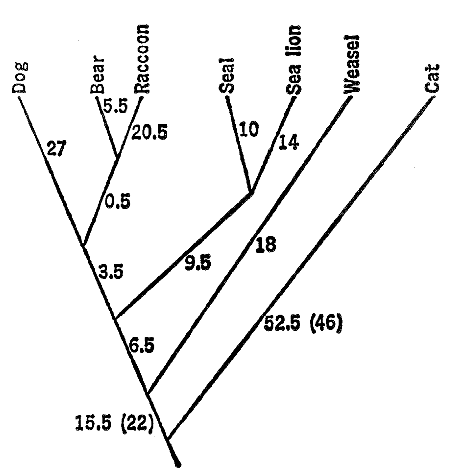


Fig. 7. Dendrogram estimating evolutionary branching relationships of seven genera of carnivores. The numbers next to each branch are estimates of the amount of evolutionary change that occurred in each line. The data are based on immunological distances treated by the Wagner distance procedure (*50*).

is well advanced. Future reduction in cost of the necessary technology will make what is feasible today, practical tomorrow. The presentation of taxonomic units in two- and three-dimensional space can be easily carried out by plotters and graphic terminals, and devices exist which will permit the inspection of the distribution of such taxonomic units from a variety of perspectives. However, the rates of automation of these procedures and the sophistication of the equipment to be employed must depend on a cost analysis of the problem to be studied and will have to be evaluated in each individual case.

Summary

There is an intimate interrelation between principles and procedures in classification, and modern work in this field has been profoundly affected by the development of electronic computers. Besides the delineation of natural systems and the achievement of economy of memory and ease of manipulation, the primary purpose of classification is the description of the structure and relationship of groups of similar objects. Successful classifications generate scientific hypotheses, although much classificatory work has applied, practical goals. The acceptance of polythetic taxa is a major conceptual advance and has directly led to classifications based on many, equally weighted characteristics. The specification of data for classification by computer will enhance objectivity but not eliminate cultural and subjective biases. Techniques of classification include cluster analysis and ordination, and numerous ways of representing classifications have been elaborated recently. By the application of graph theory to some classificatory problems it has been possible to reconstruct evolutionary branching sequences. Computer classification has been successfully applied across a broad range of disciplines.

References and Notes

1. R. A. Crowson [*Classification and Biology* (Atherton, New York, 1970)] and E. Mayr [*Principles of Systematic Biology* (McGraw-Hill, New York, 1969)] give brief historical accounts. For a detailed analysis of some of the concepts see A. J. Cain [*Ibis* **101**, 302 (1959); in *Microbial Classification*, G. C. Ainsworth and P. H. A. Sneath, Eds. (Cambridge, Univ. Press, Cambridge, England, 1962), p. 1]; also D. L. Hull [*Br. J. Philos. Sci.* **15**, 314 (1965); *ibid.* **16**, 1 (1965)].
2. E. Mayr [*Animal Species and Evolution* (Harvard Univ. Belknap Press, Cambridge, Mass, 1963), p. 17] reports all but 2 of 137 species of birds correctly differentiated by New Guinea natives, but B. Berlin, D. E. Breedlove, and P. H. Raven [*Science* **154**, 273 (1966)] found poorer correspondence between folk and scientific classifications. Raven *et al.* (*3*) note that most folk taxonomies are limited to 250 to 800 forms. For a general review see B. Berlin, D. E. Breedlove, P. H. Raven, *Principles of Tzeltal Plant Classification* (Academic Press, New York, 1974).
3. P. H. Raven, B. Berlin, D. E. Breedlove, *Science* **174**, 1210 (1971).
4. R. E. Ricklefs, *Ecology* (Chiron, Newton, Mass., 1973).
5. The Classification Society founded in 1964 in England, with a North American Branch founded in 1969, has made special efforts in this direction. See *Bull. Classif. Soc.* **1**, No. 1, to **2**, No. 4 (1965–1972) and following issues. See also general sections in Sneath and Sokal (*6*); Jardine and Sibson (*7*); J. C. Lerman, *Les Bases de la Classification Automatique* (Gauthier-Villars, Paris, 1970).
6. P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy* (Freeman, San Francisco, 1973).
7. N. Jardine and R. Sibson, *Mathematical Taxonomy* (Wiley, London, 1971).
8. J. C. Gower, in *The Assessment of Population Affinities in Man*, J. S. Weiner and J. Huizinga, Eds. (Clarendon, Oxford, 1972), p. 1; R. M. Cormack, *J. R. Stat. Soc. A* **134**, 321 (1971); R. Sibson, *J. R. Stat. Soc. B* **34**, 311 (1972).
9. R. L. Bartcher, *Kans. Geol. Surv. Comput. Contrib. No. 6* (1966); J. Rubin, *Syst. Zool.* **15**, 169 (1966).
10. The definitions given here are those of Sneath and Sokal (*6*, p. 2) and they are employed by the majority of the workers in classification.
11. G. G. Simpson, *Principles of Animal Taxonomy* (Columbia Univ. Press, New York, 1961).
12. M. Beckner, *The Biological Way of Thought* (Columbia Univ. Press, New York, 1959). Beckner's terms "monotypic" and "polytypic" were renamed "monothetic" and "polythetic" by P. H. A. Sneath [in *Microbial Classification*, G. C. Ainsworth and P. H. A. Sneath, Eds. (Cambridge Univ. Press, Cambridge, England, 1962), p. 289], and this usage has become well established.
13. Monothetic methods are frequently used for classifying associations in plant ecology [R. M. M. Crawford and D. Wishart, *J. Ecol.* **55**, 505 (1967)].
14. See discussion in chapter 8 of Sneath and Sokal (*6*).
15. See the account of a lecture by the Cambridge philosopher John Wisdom [I. J. Good, in *The Scientist Speculates: An Anthology of Half-Baked Ideas*, I. J. Good, Ed. (Heinemann, London, 1962), p. 120].
16. Two of many examples in biological taxonomy are vertebrates without erythrocytes [J. T. Ruud, *Nature (Lond.)* **173**, 848 (1954)] and higher taxa of fleas that must be defined by combinations of characters which may be lacking in a member of a taxon and which furthermore are not unique to that taxon [G. P. Holland, *Annu. Rev. Entomol.* **9**, 123 (1964)]; soils cluster well by polythetic methods and are sometimes placed into different "Great Groups" when classified monothetically on color [O. W. Bidwell and F. D. Hole, *Soil Sci.* **97**, 58 (1964); *Soil Sci. Soc. Am. Proc.* **28**, 263 (1964); J. E. Cipra, O. W. Bidwell, F. J. Rohlf, *ibid.* **34**, 281 (1970)]; in disease classification the tendency to exclude atypical cases may lead to errors [R. L. Engle, Jr., *Arch. Intern. Med.* **112**, 530 (1963)]; optical recognition of abnormal blood cells is based on polythetic criteria [N. F. Izzo and W. Coles, *Electronics* **35**, 52 (1962)]; and in archeological artifacts polythetic methods are reported to be superior to monothetic ones [J. E. Doran and F. R. Hodson, *Nature (Lond.)* **210**, 688 (1966)].
17. J. S. L. Gilmour, *Nature (Lond.)* **168**, 400 (1951); P. H. A. Sneath, *Ann. Microbiol. Enzimol.* **8**, 261 (1958). Only a few examples of generally useful classifications based on a few properties can be cited. These are usually in the physical sciences and involve some fundamenta underlying law, for example, the periodical table of elements.
18. W. B. Kendrick, *Taxon* **14**, 141 (1965); E. Mayr, *Principles of Systematic Zoology* (McGraw-Hill, New York, 1969). For an elaboration of the point of view taken in this article see P. H. A. Sneath and R. R. Sokal (*6*, p. 109).
19. K. H. L. Key, *Q. Rev. Biol.* **25**, 363 (1950).
20. J. W. Senner and R. R. Sokal, *Syst. Zool.*, in press.
21. D. Noton and L. Stark, *Science* **171**, 308 (1971); *Sci. Am.* **224** (No. 6), 34 (1971).
22. W. W. Moss, *Syst. Zool.* **20**, 309 (1971); R. R. Sokal and F. J. Rohlf, *Taxon* **19**, 305 (1970); J. D. Carroll, in *Multidimensional Scaling*, R. A. Shepard, A. K. Romney, S. B. Nerlove, Eds. (Seminar Press, New York, 1972), vol. 1, p. 105.
23. R. C. Tryon and D. E. Bailey, *Cluster Analysis* (McGraw-Hill, New York, 1970).
24. H. H. Harman, *Modern Factor Analysis* (Univ. of Chicago Press, Chicago, ed. 2, 1967).
25. J. B. Kruskal, *Psychometrika* **29**, 1, 115 (1964).
26. J. T. Curtis, *The Vegetation of Wisconsin: An Ordination of Plant Communities* (Univ. of Wisconsin Press, Madison, 1959); R. H. Whittaker, Ed., *Ordination and Classification of Communities*, vol. 5, *Handbook of Vegetation Science* (Junk, The Hague, 1972).
27. R. G. Busacker and T. L. Saaty, *Finite Graphs and Networks* (McGraw-Hill, New York, 1965).
28. W. M. Fitch and E. Margoliash, *Science* **155**, 279 (1967); *Evol. Biol.* **4**, 67 (1970).
29. A. W. F. Edwards and L. L. Cavalli-Sforza, in *Phenetic and Phylogenetic Classifications*, V. H. Heywood and J. McNeill, Eds. (Publ. No. 6, Systematics Association, London, 1964), p. 67; J. H. Camin and R. R. Sokal, *Evolution* **19**, 311 (1965); J. S. Farris, *Syst. Zool.* **19**, 83 (1970); *Am. Nat.* **106**, 645 (1972); D. G. Wallace, M. King, A. C. Wilson, *Syst. Zool.* **22**, 1 (1973).
30. R. R. Sokal and J. H. Camin, *Syst. Zool.* **14**, 176 (1965); D. H. Colless, *ibid.* **16**, 289 (1967).
31. W. Hennig, *Phylogenetic Systematics* (Univ. of Illinois Press, Urbana, 1966); G. J. Nelson, *Syst. Zool.* **19**, 373 (1970); *ibid.* **20**, 373 (1971).
32. D. L. Rosenhan, *Science* **179**, 250 (1973). See also the letters commenting on the article [*ibid.* **180**, 356 (1973)]. Although much of the exchange has little relevance to taxonomy, some of the arguments hinge directly on the definition of the meaning of disease taxa.
33. G. D. Schnell, *Syst. Zool.* **19**, 35 (1970); *ibid.*, p. 264.
34. L. Watson, W. T. Williams, G. N. Lance, *Proc. Linn. Soc. Lond.* **178**, 25 (1967)
35. R. R. Colwell, *J. Gen. Microbiol.* **58**, 207 (1969).
36. N. E. West, *Ecology* **47**, 975 (1966).
37. J. H. Rayner, *J. Soil Sci.* **17**, 79 (1966).
38. R. T. Manning and L. Watson, *J. Am. Med. Assoc.* **198**, 1180 (1966); E. S. Paykel, *Br. J. Psychiatr.* **118**, 275 (1971).
39. J. Hiernaux, in *The Assessment of Population Affinities in Man*, J. S. Wiener and J. Huizinga, Eds. (Clarendon, Oxford, 1972), p. 96; N. Jardine, *Philos. Trans. R. Soc. Lond. Ser. B* **263**, 1 (1971).
40. J. G. Jorgensen, *Salish Language and Culture* (Language Science Monographs, No. 3, Indiana Univ. Publications, Bloomington, 1969).
41. R. C. Obial, *Trans. Inst. Mining Metal. Sec. B* **79**, B175 (1970).
42. W. D. Fisher, *Clustering and Aggregation in Economics* (Johns Hopkins Press, Baltimore, Md., 1969).
43. R. E. Frank and P. E. Green, *J. Mark. Res.* **5**, 83 (1968).
44. R. B. Cattell and M. A. Coulter, *Br. J. Math. Stat. Psychol.* **19**, 237 (1966).
45. J. G. Griffith, *Mus. Helv.* **25**, 101 (1968); *J. Theol. Stud. New Ser.* **20**, 389 (1969)
46. F. Mineta, *History of the War of 1840–1842* (in Japanese, translated from the Chinese), five volumes, 1849.
47. T. T. Heine, *Simplicissimus* **30**, 181 (1925). Also in C. Schütze, Ed., *Simplicissimus Album* (Scherz, Bern, 1963).
48. The Taxocrit experiment is being conducted by R. R. Sokal and F. J. Rohlf with the assistance of R. E. Christal and W. R. Archer.
49. Data based on incidences of crimes per 100,000 people for 1970 abstracted from U.S. government sources by J. A. Hartigan, Yale University.
50. Data obtained by V. M. Sarich, analyzed by J. S. Farris, *Am. Nat.* **106**, 645 (1972).
51. Contribution No. 78 from the Program in Ecology and Evolution at the State University of New York at Stony Brook. Supported in part by NSF grant GB-35233-X00. The constructive criticisms of Professors P. H. A. Sneath and J. S. Farris are gratefully acknowledged. I also thank J. Sokal for help with Figs. 4 to 7 and Professor J. A. Hartigan for obtaining the data for Fig. 5.

References and Notes

[2] **Folk Taxonomies and Biological Classification**
Brent Berlin; Dennis E. Breedlove; Peter H. Raven
Science, New Series, Vol. 154, No. 3746. (Oct. 14, 1966), pp. 273-275.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819661014%293%3A154%3A3746%3C273%3AFTABC%3E2.0.CO%3B2-U

[2] **The Origins of Taxonomy**
Peter H. Raven; Brent Berlin; Dennis E. Breedlove
Science, New Series, Vol. 174, No. 4015. (Dec. 17, 1971), pp. 1210-1213.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819711217%293%3A174%3A4015%3C1210%3ATOOT%3E2.0.CO%3B2-2

[3] **The Origins of Taxonomy**
Peter H. Raven; Brent Berlin; Dennis E. Breedlove
Science, New Series, Vol. 174, No. 4015. (Dec. 17, 1971), pp. 1210-1213.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819711217%293%3A174%3A4015%3C1210%3ATOOT%3E2.0.CO%3B2-2

[13] **A Rapid Multivariate Method for the Detection and Classification of Groups of Ecologically Related Species**
R. M. M. Crawford; D. Wishart
The Journal of Ecology, Vol. 55, No. 2. (Jul., 1967), pp. 505-524.
Stable URL:
http://links.jstor.org/sici?sici=0022-0477%28196707%2955%3A2%3C505%3AARMMFT%3E2.0.CO%3B2-7

LINKED CITATIONS

[18] **Complexity and Dependence in Computer Taxonomy**
W. B. Kendrick
Taxon, Vol. 14, No. 5. (May, 1965), pp. 141-154.
Stable URL:
http://links.jstor.org/sici?sici=0040-0262%28196505%2914%3A5%3C141%3ACADICT%3E2.0.CO%3B2-1

[19] **A Critique on the Phase Theory of Locusts**
K. H. L. Key
The Quarterly Review of Biology, Vol. 25, No. 4. (Dec., 1950), pp. 363-407.
Stable URL:
http://links.jstor.org/sici?sici=0033-5770%28195012%2925%3A4%3C363%3AACOTPT%3E2.0.CO%3B2-X

[21] **Scanpaths in Eye Movements during Pattern Perception**
David Noton; Lawrence Stark
Science, New Series, Vol. 171, No. 3968. (Jan. 22, 1971), pp. 308-311.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819710122%293%3A171%3A3968%3C308%3ASIEMDP%3E2.0.CO%3B2-P

[22] **The Intelligent Ignoramus, an Experiment in Numerical Taxonomy**
Robert R. Sokal; F. James Rohlf
Taxon, Vol. 19, No. 3. (Jun., 1970), pp. 305-319.
Stable URL:
http://links.jstor.org/sici?sici=0040-0262%28197006%2919%3A3%3C305%3ATIIAEI%3E2.0.CO%3B2-O

[28] **Construction of Phylogenetic Trees**
Walter M. Fitch; Emanuel Margoliash
Science, New Series, Vol. 155, No. 3760. (Jan. 20, 1967), pp. 279-284.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819670120%293%3A155%3A3760%3C279%3ACOPT%3E2.0.CO%3B2-A

[29] **A Method for Deducing Branching Sequences in Phylogeny**
Joseph H. Camin; Robert R. Sokal
Evolution, Vol. 19, No. 3. (Sep., 1965), pp. 311-326.
Stable URL:
http://links.jstor.org/sici?sici=0014-3820%28196509%2919%3A3%3C311%3AAMFDBS%3E2.0.CO%3B2-3

NOTE: *The reference numbering from the original has been maintained in this citation list.*

LINKED CITATIONS

[29] **Estimating Phylogenetic Trees from Distance Matrices**
James S. Farris
The American Naturalist, Vol. 106, No. 951. (Sep. - Oct., 1972), pp. 645-668.
Stable URL:
http://links.jstor.org/sici?sici=0003-0147%28197209%2F10%29106%3A951%3C645%3AEPTFDM%3E2.0.CO%3B2-5

[32] **On Being Sane in Insane Places**
D. L. Rosenhan
Science, New Series, Vol. 179, No. 4070. (Jan. 19, 1973), pp. 250-258.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819730119%293%3A179%3A4070%3C250%3AOBSIIP%3E2.0.CO%3B2-Z

[36] **Matrix Cluster Analysis of Montane Forest Vegetation of the Oregon Cascades**
Neil E. West
Ecology, Vol. 47, No. 6. (Nov., 1966), pp. 975-980.
Stable URL:
http://links.jstor.org/sici?sici=0012-9658%28196611%2947%3A6%3C975%3AMCAOMF%3E2.0.CO%3B2-D

[50] **Estimating Phylogenetic Trees from Distance Matrices**
James S. Farris
The American Naturalist, Vol. 106, No. 951. (Sep. - Oct., 1972), pp. 645-668.
Stable URL:
http://links.jstor.org/sici?sici=0003-0147%28197209%2F10%29106%3A951%3C645%3AEPTFDM%3E2.0.CO%3B2-5

NOTE: *The reference numbering from the original has been maintained in this citation list.*